SUPPORTING MULTI DISCIPLINARY SCIENCE WITHIN NASA'S DISCIPLINE DATA SYSTEMS

Larry E. Preheim*
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, California 91109

Abstract

Many current and future NASA and non-NASA missions are focusing on multidisciplinary science. The current paradigm for data identification and effective use by the NASA science community is based on the CODMAC model proposed in 1986. As multidisciplinary investigations become more prevalent, many existing systems and the newly developing systems will have to augment the current data identification and access strategies and tools and form alliances with other data systems to provide the broad range of data required.

This paper describes the current paradigm, surveys and characterizes, within that paradigm, efforts to develop new access methods and data analysis and visualization tools, identifies additional areas of research not adequately addressed and recommends a further plan of action.

NASA's Science Environment

Missions and Science Data

There are three typical methods for obtaining scientific data from NASA.

For most of NASA's space missions the science team, formed early on, helps define the objectives and develops the instruments that are flown. The instrument may be developed by a Principal Investigator (Pi) or by a PI Team (P1 and any associated coinvestigators). They design the algorithms and code the software used to generate geophysical information from the telemetry. The science team also validates the data products. Some missions allow for an investigators "exclusive data use" period but that practice is being discouraged by most disciplines. Many fields of scientific investigation which use remotely sensed data have a well developed infrastructure. The investigators know each other and access to data is supported in a quid pro quo fashion.

There are other missions (Hubble Space Telescope, International Ultraviolet Explorer, etc.) where "facility instruments" are flown. These instruments are usually developed by a PI Team. For these instruments the majority of investigators become involved years after the instruments are designed and the products validated, They propose specific observations and are afforded time to use the facility much like an earth based telescope.

If the investigator is not directly involved with the observational system and is not associated with a colleague who is, there is still another path available. He may try to locate the data within NASA's data archives. Much of the remainder of this paper assumes the investigator pursues this third method.

---

* Member Technical Staff

Use of an Investigator's Time

In 1989 the Science Advisory Panel for Eos Data and I nformation provided an estimate of how much effort glees into a typical research paper[1]. That information is repeated here as Table 1. According to the estimate, close to 20% of a researcher's time is spent in dealing with the data system(s) to locate and obtain the input data necessary (steps 1, 2 and 5) and about 70% is spent sorting and understanding the data (steps 3,4,6 and 7).

The sorting and analyzing activities are often supported by tools developed by and familiar to the researcher. In addition, NASA is investing heavily in development of supplemental tools to support the data analysis steps. Data visualization is a high priority research activity within the Information Systems Branch, NASA Code SMI. In general, less research emphasis is placed on locating and ordering the data.

Table 1 Production of a Research Paper

Step	Fraction (percent)	Time (days)
1. Deciding on data to use	5	3
2. Ordering and receiving data	10	6
3. Preliminary sorting	15	9
4. Preliminary Relations	10	6
5. Ordering supplemental data	5	3
6. Reselection and resorting	25	15
7. Refinement of relations	20	12
8. Writing and publishing	10	6
Total	100	60

CODMAC Data Management Model

In 1986 a National Research Council report[2], generally referred to as the CODMAC (Committee on Data Management and Computation) Report, established a specific model for management of NASA's space science data. That model depicted in below is still the standard throughout NASA.

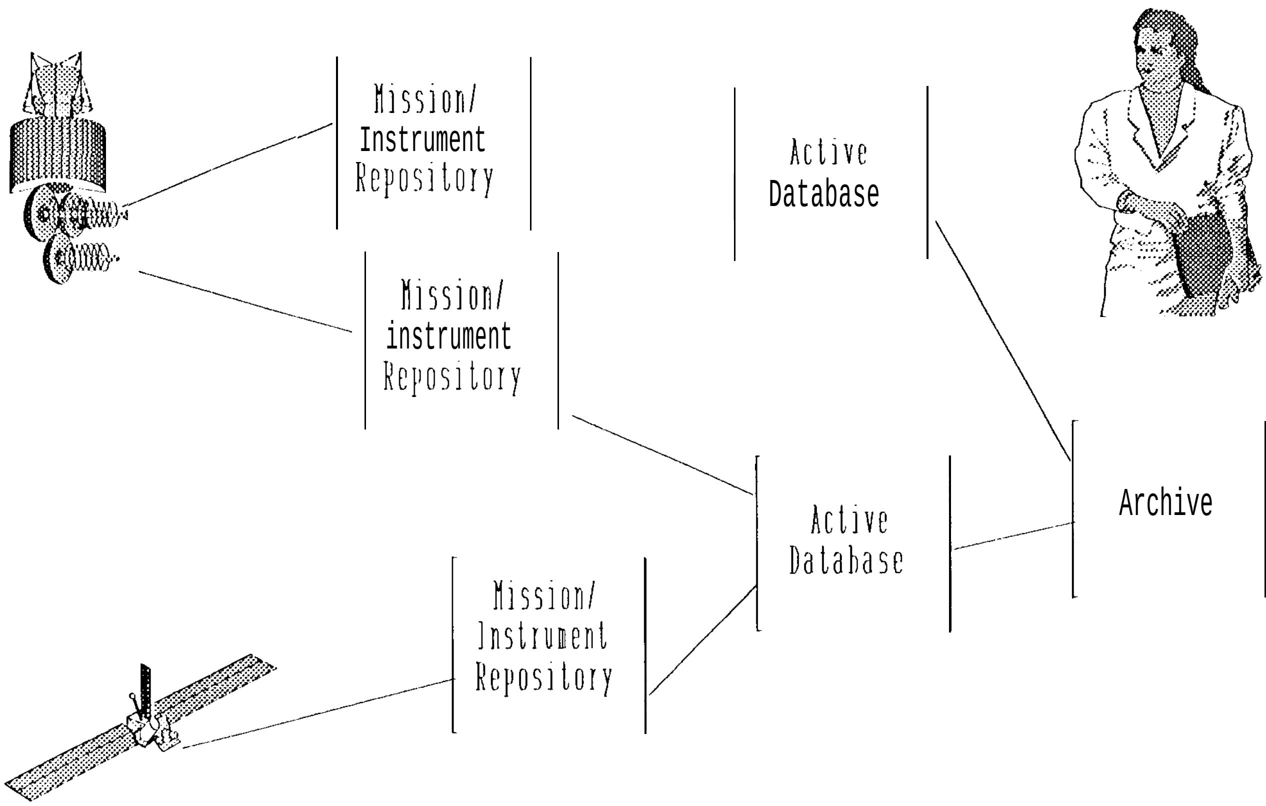

Figure 1 CODMAC Data Management Model

Data originates at the instrument. It is captured and passed on to one or more Data Repositories. Data Repositories are facilities that are temporary buffers for new data, usually existing only as long as the mission producing the data. There, initial data processing, validation and correlation take place. Then the data are passed on to Active Database Sites. Active Database Sites are locations which house data actively used in ongoing research and are generally under the direct control of and housed with the science community using them. Eventually the data finds its way into a discipline specific archive. Active Database Sites relay recall or continue to maintain data previously passed to an Archive. This is likely done when the older data generates new interest because a new data set can be correlated with it in some meaningful way; there is now some interest in reviving the Surveyor lunar data. Archives are stable databases that ensure long-term access to the data by the general science community. All three Space Science Data Management Units, as the report terms them, have heavy science user community involvement and management oversight, thus guaranteeing their responsiveness to the user community.

One of the biggest challenges facing the secondary and interdisciplinary investigators using EOSDIS or the or any discipline data system will be how to find and use the few megabytes of interesting data among the many pitabytes being managed by the Information Systems. This was stated in the CODMAC Report[3] as "A major impediment in space sciences research is the lack of information about what data sets exist, what their characteristics are, and how to obtain calibrated versions of the data."

Data Search Model

The data search model generally assumed to provide the most help to investigators involves a layered approach to data information abstraction (metadata). This hierarchy is depicted as figure 2 below. The CODMAC Report supported this approach and recommended a strong role for the NASA Master Directory (NMD). At the top layer is a master directory. The NASA Master Directory is chartered to provide[4] "a starting point for a NASA-funded researcher or other scientists to do a computer-aided search for earth and space science data of interest"

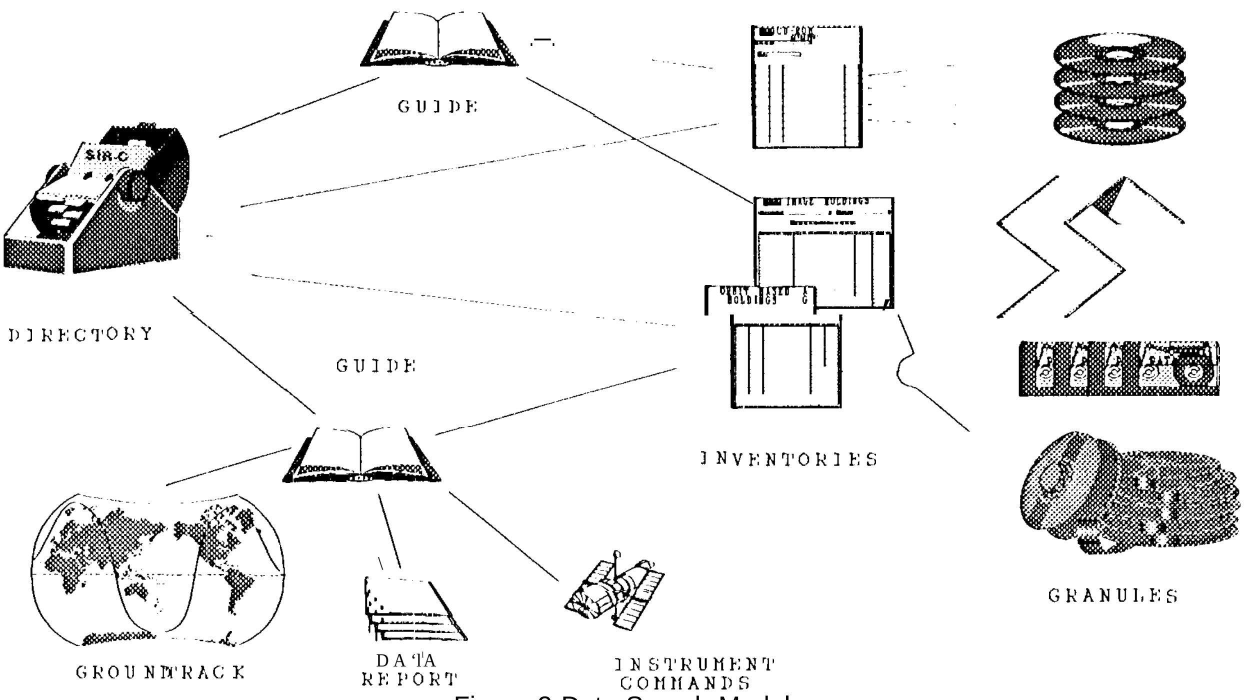

Figure 2 Data Search Model

The next layer down, usually at a discipline specific data system (either an Active Database Site or an Archive, but generally dealing with one NASA Science Division - Life Sciences, Earth Sciences, Microgravity, Planetary Science, Space Physics or Astrophysics), is generally termed a Guide. Additional information about the data holdings and instrument calibrations and campaign descriptions reside with this collection of information. A mechanism called "context passing" has been created to preserve the data selection/search criteria already entered as the investigator is transferred between the NMD and the Archive systems.

The bottom layer of the metadata hierarchy is the description of the data granule itself. A granule is simply the unit at which the data is locally managed (reel of tape, day or revolution organized file, single image or swath file, etc.) Data associated with a granule is not very standard. However as part of the recent data census conducted by NASA, a template of items about the data granule was created. This template may form the basis for standard data interchange format for granules similar to the standard for exchange of data at the master directory level. No work in this area is currently proposed.

Investigators familiar with the data or having experience with either an Active Database Site or an Archive are expected to go directly to the inventory to find their information and then perhaps to the Guide to verify some detail. Investigators with no previous experience are expected to work their way down from the Directory (i.e. the NMD) and may go to the inventory next or browse the guide first. The NASA Master Directory is a mature, well understood entity with its own de facto standards. The inventory is specific to the data system and to the granule managed but widely implemented. The middle ground, the Guide, has yet to be defined and generally accepted.

Multidisciplinary Science

Current Status

NASA has had multi-disciplinary data systems for quite some time. The Space Sciences Data Operations Office (SSDOO) at Goddard Space Flight Center, (organizational successor to NSSDC - National Space Science Data Center), PDS (Planetary Data System) and PLDS (Pilot Land Data System) are examples of data systems supporting more than one specific science discipline. Indeed, with its role as the entry into NASA's data holdings the NMD supports all five of NASA's science divisions. The emphasis, however, appears to be on supporting several disciplines rather than an integrated multidisciplinary field of study.

The requirements for support of multidisciplinary science are founded, in large measure on the scope of problems addressed by the Global Change Research Program (GCRP). EOS, based in the Earth Sciences Division, as NASA's main contribution to the GCRP compels a new approach to making data available.

In addition to its instrument PIs and PI Teams, the EOS Project has selected twenty nine multidisciplinary Principal Investigator (Pi) or PI Teams to participate in the program. The EosDIS Science Advisory Panel for Eos Data and Information has representatives from the multidisciplinary investigators. However, not one of the supported multidisciplinary PIs can get all the data he needs from EOS; each needs one or more data products from an additional source. Further, of the data products available from EOS, not a single Pi's needs are scheduled to be completely met within this century[5]; all require at least one source of data to be flown on the second EOS platform.

A work item under development for the EOSDIS Phase O implementation concerns providing a standard method and interface for ordering granules of data*. This could be done in an automated way for both same-day and next day action. The service could sort the requests by archive and optimize work lists for the information systems as well as providing user access validation and central billing services for the system. However this service is EOSDIS specific and does not cover the NASA discipline or other Agency data systems.

Ongoing Research

The most concentrated program of ongoing research is the Applied Information Systems Research Program funded by the Information Systems Branch of the Flight Systems Division, NASA Code SMI. A categorization of the research they are funding is given as table 2 below. Various functional areas are further described in the following paragraphs.

Data Format Conversion: NASA has, in the past supported the development of many "standards" for data formatting. Examples of existing standards include FITS (Flexible Image Transport System, an astrophysics standard), CDF (Common Data Format, a NSSDC standard), SFDUs (Standard Formatted Data Units, a CCSDS standard) and several more. If you consider commercially supported image standards, there are more than twenty available in the market place today. This proliferation of data "standards" requires the investigator to be able to convert between what the data Archive offers and what his/her analysis software expects. One of the more useful data conversion tools as far as current flavors of NASA data is concerned is the Data Hub reported in the previous paper.

In June of 1992 the NASA /OSSA Office of Standards and Technology (NOST) held an invitational workshop to discuss data formats. Representatives of ten popular format camps were invited. NOST believes that as a result of the

Table 2 Applied Information Systems Research Program

Title	AI Analysis	Data Compression	Data Visualization	Data Access
Multivariate Statistical Analysis Software Technologies	X			
Multi-layer Holographic Bifurcative Neural Network	x			
Expert Data Reduction Assistant	X			
Scientists' Intelligent Graphical Modeling Assistant	X			
Parallel Algorithms for Data Compression		X		
Performance and Scabability of Client Serve Database		X		
EOSDIS Testbed System		X		
GIS for Fusion and Analysis of High Resolution Data	X	X	X	
Analysis and Displlay System for Large Geophysical Data sets	X		X	
Grid Analysis and Display System			X	
NCAR Interactive			X	
A Distributed System for Visualization and Analysis			X	
Planetary PC-McIDAS			X	
Experimenter's Laboratory for Visualized Interactive Sci.			X	
Space Data Analysis and Visualization System			X	
NASA Supported Advanced visualization Techniques			X	
LinkWinds			X	
DataHub				X

meeting it has identified some issues that should be considered by those choosing formats. However the workshop report has yet to be published.

Visualization: The effective visualization of science data has been an active topic for discussion and research among NASA's computer scientists and data system development folks for some time. The *Information Systems Newsletter* (the quarterly publication of the Information Systems Branch of NASA's Office of Space Science and Applications) has featured an article on the subject in about every other issue.

A recent review' of the applicability of visualization to EOS reported the following conclusion: "Although our present suite of visualization techniques is impressive and powerful, the application software for putting these capabilities into the hands of the scientist is, at present, inadequate."

Browse Data Compression; NASA has recognized the value to investigators of using browse data to identifying useful (image) data for further analysis. Several discipline data systems (Active Databases) have developed and used browse software as part of their user data selection interfaces. However, it is only just beginning to generate requirements and standards for such functions

Data Element Dictionaries: One of the noticeable trends among discipline data systems is to provide data format and content documentation with the distributed data in the form of a data element dictionary (DED). For the PDS this metadata is available online but is not normally distributed with the science data. A DED provides for a description of each field including rules of formation, units (ergs, dynes, lumens/cc, etc.), discrete lists of contents and more. A DED can be used both to interpret the data and to validate the contents of a given record. A DED could also be used as a basis for "automatic translation" of data from one format to another (ergs/square centimeter to millibars) which would be more helpful and familiar to a data user. The Consultive Committee for Space Data Systems (CCSDS) Panel 2 has begun work on a standard for DEDs.

SPICE: For many of NASA's past missions the PI or the Project's PI Representative (for the larger missions) was the holder of most of the knowledge about an instrument and the data gathering sequences. Items like "every time we go into occultation channel B hiccups" and "there is no data from . . . because..." were passed down by word of mouth or written in a lab notebook and not documented as part of the formally archived history of the data. Now, for some data sets this information is available. NASA is funding the development of SPICE[9] (Spacecraft ephemeris, Planet ephemeris and constants, Instrument descriptions, Camera pointing and Events), a standard set of ancillary data files and software (called S, P, ... Kernels for each information type) for their manipulation. This system will provide investigators with the indicated types of information in a fashion which will allow more complete understanding of how the observations were taken.

The SPICE system has been distributed and used internationally. At this time no Archive distributes spice data with its science product although PDS does have SPICE data for Voyager available[10]. It is possible that assembling SPICE data for some of the more interesting past missions will be impossible or economically infeasible. Mars Observer is using the Events Kernel (E-Kernel) and would have liked to reference the Viking Events data in preparing for their own mission operations[11].

Additional Efforts Needed

Automatic Metadata Loading: EosDIS[12] is tracking 3,704 input and output data products to be used in daily processing. They are expecting to produce 1.2 gigabytes of data per day (GB/d). Almost half of that total is expected to be managed by the Goddard DAAC (Distributed Active Archive Center). EosDis expects to move around 25 GB/d of level 1 B (engineering units) data and 10 GB/d of level 2/3 (geophysical) data. Keeping track of the state and location of this data in some form more useful than a simple inventory entry is going to be impossible without intelligent tools to help. JPL is experimenting with intelligent assistants to help in setting up, monitoring, fault detection and recovery, and post processing reassignment of subsystems to support Deep Space Network (DSN) tracking passes. Such technology may eventually prove applicable to managing the data reduction of large volumes of EOS data. The successful solution must also generate and post the necessary updates to the EOSDIS Directory, Guides and Inventories.

Knowbots: Intelligent assistants who know about the way an investigator works and what types of information interest him or her would be a useful asset. An electronic version of such an assistant called a Knowbot was described in ***Scientific American***[13] in 1990. At about the same time, a system architecture and an architectural element called "Mediators" was described[14] by Gio Wiederhold. The purpose of both these two constructs would be to provide an active agent on behalf of information seekers which could act both as a data sleuth and data format converter. Given the size and complexity of NASA's anticipated data holdings, Knowbots should enable discovery of useful information without Herculean) efforts by those not steeped in the ontology of the data systems.

Better Coordination with other science efforts is needed. If none of the multidisciplinary PIs selected by NASA can complete their research with data from a data source the size of EOS, it is likely other agency science programs have similar problems. One high level review[15] of EOS put it this way: "The EOSDIS program must be structured and managed to facilitate interactions with the other agencies involved in the G.S. Global Change Research Program so that existing data and future data collected by NASA and by other national and international organizations -- using research and operational satellites as well as in situ sources -- are available to all global change research scientists."

Long Term Archiving is also an issue, One aspect of some of the multidisciplinary research interests under consideration is the long time scale of the phenomena, in some cases years. Yet some of NASA's past missions data currently residing in storage (on the assumption that it may be valuable again someday) is economically unusable. Some of this data is on 200bpi 7-track tape and for some of the data, the field definitions have been lost. The previously mentioned review[16] of EOSDIS says: "Long-term archiving of EOS data is an issue that has not been addressed. Long-term commitment to maintaining data collected as part of EOSDIS is a critical component of the U.S Global Change Research Program." Application of emerging data format and labeling standards as well as media format standard needs to be examined.

Data Location: Although NASA has chartered multidisciplinary Pi's to work with EOSDIS and to help it develop products for that community, there is still a larger group who would benefit from easy access to the data. Catalogs (Directory/Guide/inventory) are the keys to finding and making effective use of the data. There is wide spread belief that making useful catalogs will continue to be a problem: "'There is and will continue to be diversity in the form and content of discipline data system catalogs that discourages multi-disciplinary searches"[17] and "The next ten fold increase in the amount of information in the databases will divide the world into haves and have-nets, unless each of us connects to that information and learns how to sift through it for the parts we need."[18]

NASA can not put the onus on the multidisciplinary scientists to understand and use existing data storage paradigms for location of information. NASA must adopt for all of its archives an interface flexible enough to allow investigators to search for data within the context of their own information framework.

ESA (the European Space Agency) is performing some interesting studies in this area. They have approached the problem from the tack "interface design should be embedded in whole-system design, rather than tacked onto some unrelated, predefine structure."[19] They have worked out a set of semantic networks, one for each discipline they support which is a set of nodes and arcs indicating the data concepts and the relations between them. That is, the semantic nets map the way data relationships are presented to the user. The are planning to allow users to edit their own semantic nets to customize their logical data interface. In the spring of 1993 workshops were held at ESA/ESRIN to introduce users to the new system. User evaluations have not yet appeared in print,

Conclusions

NASA is working hard at supporting use of its vast holdings (and holdings-to-be) of data by the multidisciplinary science community. The primary focus seems to be on visualization with assisted analysis coming in second place. Less well addressed are issues with how the user locates and obtains the data. This is lack particularly crucial in the anticipated multi-mission, multi-agency environment necessary to supply data to multidisciplinary investigators. NASA must lead and fund an effort to develop a new paradigm for data location and ordering which is flexible enough to adapt to the data semantics familiar to the investigator whatever they may be.

The work described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

1. "initial Scientific Assessment of the Eos Data and Information System (EosDIS)", Science Advisory Panel for Eos Data and Information, NASA, Eos-89-1, page 30, 1989

2. "Issues and Recommendations Associated with Distributed Computation anti Data Management Systems for the Space Sciences", National Academy Press, 1986

3. "Issues and Recommendations Associated with Distributed Computation and Data Management Systems for the Space Sciences", National Academy Press, 1986, pg 7

4. "Catalog Ilnteroperability Master Directory Project Plan", Draft Version 3.2, April 1989

5. "Earth Observation System Output Products and Input Requirements", Version 2.0, Science Processing Support Office (SPSO), GSFC, Vol. 2, page 4, August 1992

6. McDonald, Ken, One StopShopping for Locating Earth Science Data, Information Systems Newsletter, p 4, March 1991

7. Botts, Michael, "The State of Scientific Visualization with Regard to the NASA EOS Mission to Planet Earth", Earth Systems Science Laboratory, University of Alabama in Huntsville, p. 21, August 1992

8. Tilton, Jim, "NASA's Data Compression Needs Discussed at Workshop", Information Systems Newsletter, p. 7, July 1991

9. Acton, Chuck and Underwood, Ian, "What is SPICE", NASA OSSA Information Systems Newsletter, pp 36-41, February 1990

10. Lynch, John, NAIF Participates in Voyager SEDR Regeneration, Information Systems Newsletter, pp 6-7, February 1992

11. Acton. Chuck, NAIF Delivers E-kernel Subsystem to' Mars Observer & MARS 94/96, Science Information Systems Newsletter, p. 19, April 1993

12. Earth Observation System Output Products and Input Requirements, Version 2.0, Science Processing Support Office (SPSO), GSFC, Vol. 1, August 1992

13. Wright, K., The Road to the Global Village, Scientific American, pp 84-94, March 1990

14. Wiederhold, Gio, Future Architectures for Information Processing Systems, PARBASE-90 International Conference on Databases, Parallel Architectures, and Their Applications, IEEE Society Press, pp 160-176, May 1990

15. National Research Council Panel to Review EOSDIS Plans, Interim Report, National Research Council, page 2, April 1992

16. National Research Council Panel to Review EOSDIS Plans, Interim Report, National Research Council, page 9, April 1992

17. Davis, Randal, NASA's Science Discipline Data Systems: A Functional Comparison, LASP, University of Colorado at Boulder, presented at the international Space Year Conference on Earth and Space Science Information Systems, Pasadena CA, February 1992

18. Gilbert, Walter, Towards a paradigm shift in biology, Nature, Vol 349, page 99, 10 January 1991.

19. Belkin, N., etal, User Interfaces for Information Systems, ESA Journal, Vol 14, pp 325-342, 1990